

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization International Bureau



(43) International Publication Date  
6 May 2005 (06.05.2005)

PCT

(10) International Publication Number  
WO 2005/041063 A1

(51) International Patent Classification<sup>7</sup>: G06F 17/30

(21) International Application Number:

PCT/GB2004/004028

(22) International Filing Date:  
22 September 2004 (22.09.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
0322899.6 30 September 2003 (30.09.2003) GB  
0328890.9 12 December 2003 (12.12.2003) GB

(71) Applicant (for all designated States except US): BRITISH TELECOMMUNICATIONS PUBLIC LIMITED COMPANY [GB/GB]; 81 Newgate Street, London, Greater London EC1A 7AJ (GB).

(72) Inventors; and

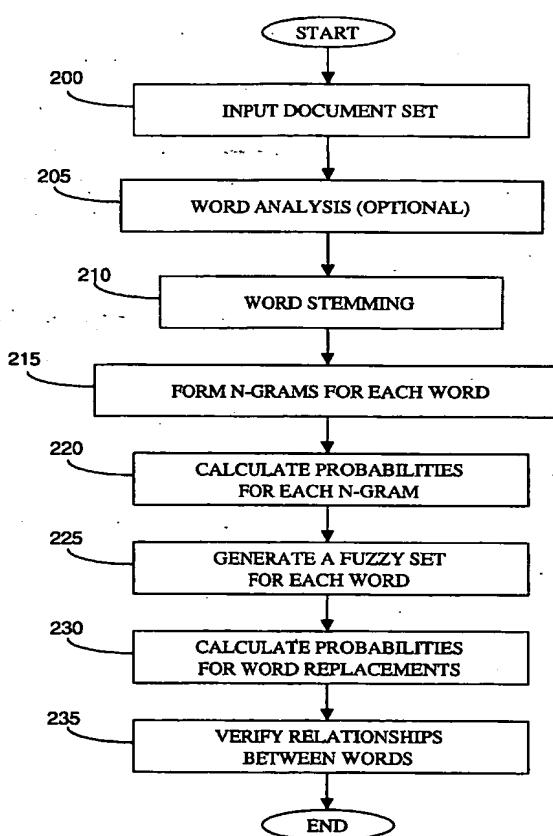
(75) Inventors/Applicants (for US only): ASSADIAN, Behrad [GB/GB]; 2 Knights Lane, Kesgrave, Ipswich, Suffolk IP5 2FS (GB). AZVINE, Behnam [GB/GB]; 6 Dodson Vale, Kesgrave, Ipswich, Suffolk IP5 2GT (GB). MARTIN, Trevor, Philip [GB/GB]; 86 High Street, Cam, Gloucestershire GL11 5LH (GB).

(74) Agent: GEFFEN, Nigel, Paul; BT Group Legal Intellectual Property Department, PPC5A, BT Centre, 81 Newgate Street, London, Greater London EC1A 7AJ (GB).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM,

[Continued on next page]

(54) Title: INFORMATION RETRIEVAL



(57) Abstract: A method and apparatus are provided for generating, from an input set of documents, a word replaceability matrix defining semantic similarity between words occurring in the input document set. For each word, distinct word sequences of predetermined length are identified from the documents of the set, each word sequence being indicative of the context in which the word was used and, according to the relative frequency of occurrence of the identified word sequences for the word, fuzzy sets are generated for each word comprising membership values for corresponding groups of word sequences. For each pair of words occurring in the document set, their respective fuzzy sets are used to calculate the probability that the first word of a pair is semantically suitable as a replacement for the second word of the pair, these probabilities being collated to form a word similarity matrix for use in an improved method of determining document similarity and in information retrieval.

WO 2005/041063 A1



TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM,  
ZW.

SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ,  
GW, ML, MR, NE, SN, TD, TG).

**(84) Designated States (unless otherwise indicated, for every kind of regional protection available):** ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI,

**Published:**

— with international search report

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*